

(19)日本国特許庁 (J P)

(12) 公 開 特 許 公 報 (A)

(11)特許出願公開番号
特開2002-163140
(P2002-163140A)

(43)公開日 平成14年6月7日(2002.6.7)

(51)IntCl. ⁷	識別記号	F I	テ-マコ-ト*(参考)
G 0 6 F 12/00	5 4 5	G 0 6 F 12/00	5 4 5 B 5 B 0 1 4
13/10	3 4 0	13/10	3 4 0 A 5 B 0 8 2

審査請求 未請求 請求項の数5 O L (全 16 頁)

(21)出願番号 特願2000-359810(P2000-359810)

(22)出願日 平成12年11月27日(2000.11.27)

(71)出願人 000005223

富士通株式会社

神奈川県川崎市中原区上小田中4丁目1番
1号

(72)発明者 大江 和一

神奈川県川崎市中原区上小田中4丁目1番
1号 富士通株式会社内

(72)発明者 西川 克彦

神奈川県川崎市中原区上小田中4丁目1番
1号 富士通株式会社内

(74)代理人 100092978

弁理士 真田 有

Fターム(参考) 5B014 EB04 FA05

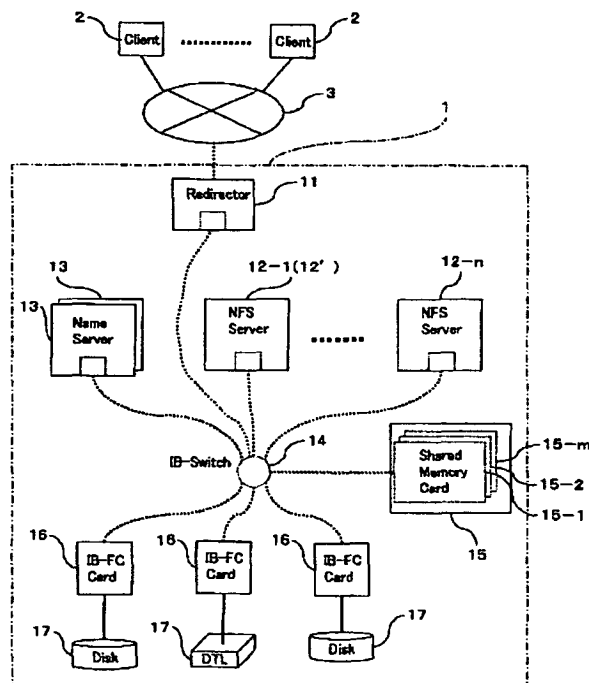
5B082 AA01 FA16 HA01 HA05 HA08

(54)【発明の名称】 ストレージシステム

(57)【要約】

【課題】 ネットワークの帯域拡大に対して低コストで十分に対応できるスケーラビリティをもったストレージシステムを提供することを目的とする。

【解決手段】 ファイルデータを記憶しうる記憶装置17と、ファイルデータに関するリクエストに応じたファイル処理を記憶装置17に対して行なう複数のファイルサーバ12-1~12-nと、外部ネットワーク3を介してクライアント2から受信されるファイルリクエストのファイルサーバ12-i(i=1~n)への転送処理と、そのファイルリクエストに対するクライアント2への応答処理とを管理するファイルサーバ管理ノード11と、記憶装置17、ファイルサーバ12-i及びファイルサーバ管理ノード11を通信可能に相互接続する内部ネットワーク14とをそなえるように構成する。



【特許請求の範囲】

【請求項1】 ファイルデータを記憶しうる記憶装置

と、
該ファイルデータに関するリクエストに応じたファイル処理を該記憶装置に対して行なう複数のファイルサーバと、
外部ネットワークを介してクライアントから受信されるリクエストの該ファイルサーバへの転送処理と、該リクエストに対する該クライアントへの応答処理とを一元管理するファイルサーバ管理ノードと、
該記憶装置、該ファイルサーバ及び該ファイルサーバ管理ノードを通信可能に相互接続する内部ネットワークとをそなえて構成されたことを特徴とする、ストレージシステム。

【請求項2】 該内部ネットワークに、該ファイルサーバが扱うファイルデータ名を一元管理するネームサーバが接続されていることを特徴とする、請求項1記載のストレージシステム。

【請求項3】 該内部ネットワークに、該ファイルサーバ管理ノード及び該ファイルサーバがアクセス可能な共有メモリが接続されていることを特徴とする、請求項1記載のストレージシステム。

【請求項4】 該内部ネットワークに、該ファイルサーバ管理ノード、該ファイルサーバ及び該ネームサーバがアクセス可能な共有メモリが接続されていることを特徴とする、請求項2記載のストレージシステム。

【請求項5】 該ファイルサーバ管理ノードが、該リクエストの内容を解析するリクエスト解析部と、該リクエスト解析部の解析結果に応じて該リクエストを特定のファイルサーバに転送するリクエスト転送部とをそなえていることを特徴とする、請求項1～4のいずれか1項に記載のストレージシステム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、ストレージシステムに関し、所望のネットワークに接続されることで複数のクライアントでファイルデータの共有を可能にするストレージシステムに関する。

【0002】

【従来の技術】ネットワーク上での複数ノード（クライアント）間のファイルデータの共有（以下、単に、「ファイル共有」という）を実現する従来の手法としては、例えば図16に模式的に示すように、ネットワークファイルシステム（NFS：Network File System）を利用してLAN（Local Area Network）などの所望のネットワーク100上にファイルサーバ200を構築し、このファイルサーバ200にSCSI（Small Computer System Interface；一般に「スカジー」と呼ばれる）などのインタフェース300を介して二次記憶装置400を接続して、この二次記憶装置400において複数クライ

アント500間のファイル共有を実現する方法が良く知られている。

【0003】しかしながら、この方法では、次のような課題がある。

①ファイルサーバを構築・維持（保守）するのに専門的なスキルが必要である。

②ファイルサーバの拡張（容量、アクセス性能）が容易でない。拡張できてもファイルサーバが複数に分かれたりしてしまうなどで維持コストが増大してしまう。

【0004】③故障時にそなえたシステム構築・維持（保守）に専門的なスキルが必要であり、また、そのための費用もかかる。

これらの課題を解決する方法として、近年、NAS（Network Attached Storage）が提案されている。このNASは、上記のファイルサーバ200及び二次記憶装置400から成る部分（図16中の破線枠参照）が予め1つのストレージシステムとして構築されたものに相当し、ネットワーク100に接続して簡単な設定を行なうだけで、ファイル共有が実現できるシステムで、システムの構築・維持（保守）に専門的なスキルは必要ない。

【0005】

【発明が解決しようとする課題】しかしながら、このようなNASにおいても、現在急速に進んでいるLANの帯域拡大（現状で1Gbps、数年後には10Gbps程度）に低コストで十分に対応できるスケーラビリティが得られていないという課題は残っている。即ち、接続先のネットワークの帯域拡大に対応しようとすると、NASにおいても、単に、内部のファイルサーバ及び二次記憶装置を増設することになり、この結果、ファイルサーバが複数に分かれてしまい、それぞれが管理する二次記憶装置も分割されてしまうことになる。

【0006】つまり、上記のファイルサーバ200及び二次記憶装置400を並列（独立）して設けることになる。このため、結局、ファイルサーバの維持（保守）を個々に行なう必要があり、維持コストが増大してしまう。本発明は、このような課題に鑑み創案されたもので、ネットワークの帯域拡大に対して低コストで十分に対応できるスケーラビリティをもったストレージシステムを提供することを目的とする。

【0007】

【課題を解決するための手段】上記の目的を達成するために、本発明のストレージシステム（請求項1）は、ファイルデータを記憶しうる記憶装置と、この記憶装置に対しリクエストに応じたファイル処理を行なう複数のファイルサーバと、外部ネットワークを介してクライアントから受信されるリクエストの上記ファイルサーバへの転送処理と、そのリクエストに対するクライアントへの応答処理とを一元管理するファイルサーバ管理ノードと、上記の記憶装置、ファイルサーバ及びファイルサーバ管理ノードを通信可能に相互接続する内部ネットワー

クとをそなえて構成されたことを特徴としている。

【0008】ここで、上記の内部ネットワークには、上記のファイルサーバが扱うファイルデータ名を一元管理するネームサーバが接続されていてもよい（請求項2）、上記のファイルサーバ管理ノード及びファイルサーバがアクセス可能な共有メモリが接続されていてもよい（請求項3）。なお、上記のネームサーバが存在する場合は、上記のファイルサーバ管理ノード及びファイルサーバに加えて、このネームサーバも上記の共有メモリにアクセスが可能である（請求項4）。

【0009】また、上記のファイルサーバ管理ノードは、例えば、上記のリクエストの内容を解析するリクエスト解析部と、このリクエスト解析部の解析結果に応じて該リクエストを特定のファイルサーバに転送するリクエスト転送部とをそなえて構成されるのが好ましい（請求項5）。

【0010】

【発明の実施の形態】以下、図面を参照して本発明の実施の形態を説明する。

（A）一実施形態の説明

図1は本発明の一実施形態としてのストレージシステムの構成（ストレージアーキテクチャ）を示すブロック図で、この図1に示すストレージシステム1（以下、単に「システム1」ともいう）は、外部ネットワーク（例えば、ギガビットイーサネット（登録商標））3に接続された複数のクライアント2間でファイル共有を実現するためのものであって、リダイレクタ（Redirector）11、複数のNFSサーバ（ファイルサーバ）12-1～12-n、ネームサーバ13、共有メモリ（Shared Memory）15、I B-F Cカード16、二次記憶装置17（ディスク装置やテープ装置（DTL）など）をそなえており、これらの各コンポーネント11、12-i、13、15、16、17が例えば4～10Gbps（ギガビット/秒）程度の高速（内部）ネットワーク〔インフィニバンド（Infiniband）〕スイッチ14を介して相互に接続された構成になっている。

【0011】このようにシステム内部の各コンポーネント（リダイレクタ11、NFSサーバ12-i、ネームサーバ13、二次記憶装置17など）を内部ネットワーク14で接続する形態（クラスタリング）をとることで、必要に応じてNFSサーバ12-iや二次記憶装置17を内部ネットワーク14に接続すれば簡単に増設することができ、外部ネットワーク3の通信速度などに応じた拡張性（容量、アクセス性能など）が大幅に向上する。

【0012】ここで、上記のリダイレクタ（ファイルサーバ管理ノード）11は、外部ネットワーク3を介して任意のクライアント2から受信される各種リクエストメッセージ（以下、単に「リクエスト」という）のNFSサーバ12-i（i=1～n）への転送処理と、そのリ

クエストに対するリクエスト元のクライアント2への応答処理とを一元（集中）管理するためのものである。つまり、本リダイレクタ11の存在により、上述のごとくNFSサーバ12-iや二次記憶装置17を増設したとしても、従来のようにそれらを個々に保守する必要は無いのである。

【0013】なお、上記の「リクエスト」は、二次記憶装置17に記憶されたファイルデータ（以下、単に「ファイル」ともいう）に関する要求を意味し、例えば、ファイルデータの実体に対するファイル操作（書き込み／更新／読み出しなど）要求（ファイルアクセスリクエスト）や、それ以外のファイル名参照などのメタ情報アクセスなどがある。また、各クライアント2からは、基本的に、このリダイレクタ11に付与されたIP（Internet Protocol）アドレスのみが参照できる。つまり、各クライアント2からは、本システム1が1つのファイルサーバとして見えるようになっている。

【0014】そして、上述したような機能を実現すべく、本リダイレクタ11には、例えば図2に示すように、外部ネットワーク3用のインタフェースを装備するギガビットイーサネットカード11a、システム内部（内部ネットワーク）用のインタフェースを装備するインフィニバンドカード11b、これらの各カード11a、11bをはじめとして自身（リダイレクタ11）の動作を集中制御するためのネットワークプロセッサ11c、このネットワークプロセッサ11cが動作する上で必要なソフトウェア（プログラム）や各種データを記憶するためのメモリ（主記憶部）11dなどが実装される。

【0015】なお、ネットワークプロセッサ11cは、これらのコンポーネント11a～11dとPCI（Peripheral Component Interconnect）バス11eを介して相互通信可能に接続されている。ここで、上記のネットワークプロセッサ11cは、内部ネットワーク14と外部ネットワーク3との間で送受されるリクエストやそのリクエストに対するリプライ（応答）メッセージなどの送受（プロトコル変換なども含む）や、クライアント2から受信されるリクエスト（プロトコル）の解析、その解析結果に基づいたアクセスファイル名の決定、受信リクエストの転送先NFSサーバ12-iの決定などを行なうことができるもので、本実施形態では、次のような制御も可能になっている。

【0016】即ち、クライアント2からのリクエストを解析し、各ファイルサーバ12-iに均一に（例えば、負荷の軽いNFSサーバ12-iに）リクエストが転送されるように制御したり、同一ファイルについてのリクエストは同じNFSサーバ12-iに割り当てて、NFSサーバ12-i間でファイルアクセス競合が起きないように制御したりすることができるようになっている。

【0017】このため、本ネットワークプロセッサ11

cには、その要部の機能に着目すると、次のような機能が実装されている。

(1) クライアント2からのリクエストの内容を解析するリクエスト解析部111としての機能

(2) このリクエスト解析部111の解析結果に応じて受信リクエストを特定のNFSサーバ12-iに転送するリクエスト転送部112としての機能

(3) リクエスト転送部112による過去のリクエストの転送履歴を(例えば、メモリ11dに)記録する転送履歴記録部113としての機能

(4) NFSサーバ負荷監視デーモン(daemon)115をバックグラウンドで実行することにより、各NFSサーバ12-iの負荷状態を定期的に監視する負荷監視部114としての機能

これにより、上記のリクエスト転送部112は、上記の転送履歴記録部113による転送履歴に基づいて同一ファイル名のファイルに対するリクエストを同一NFSサーバ12-iに転送したり、NFSサーバ負荷監視デーモン115(負荷監視部114)による負荷監視結果に基づいて負荷の低いNFSサーバ12-iに受信リクエストを転送したりすることが可能になる。

【0018】従って、処理速度が大幅に向上するとともに、各NFSサーバ12-iに均一にリクエストを転送できるので、一部のNFSサーバ12-iに負荷が集中して障害を引き起こしてしまうようなことを確実に防止することができ、信頼性が大幅に向上する。なお、本ネットワークプロセッサ11cは、例えば、メタ情報アクセスに対するリプライメッセージを内部のキャッシュメモリ11f(メモリ11dでもよい)にキャッシュしておき(図3参照)、クライアント2からメタ情報アクセスについてのリクエストを受けると、まず、キャッシュメモリ11f(あるいは、メモリ11d)をチェックして、ヒットすればリプライメッセージを作成してそのまま(NFSサーバ12-iやネームサーバ13に転送することなく)クライアント2側に返す(図4参照)ようにすることもできる。

【0019】この仕組みは、メタ情報のみならずファイルデータについても適用できる。ただし、この場合、全てのファイルデータをキャッシュするようにすると必要なメモリ容量が増大するので、ネットワークプロセッサ11cでアクセス頻度の高いファイルデータを選び出し、そのファイルデータのみをキャッシュメモリ11f(あるいは、メモリ11d)にキャッシュするようにした方がよい。

【0020】つまり、上記のキャッシュメモリ11f(あるいは、メモリ11d)は、リクエスト頻度の高い特定のファイルデータについてのクライアント2へのリプライメッセージをキャッシュしておくキャッシュ部としての機能も果たし、この場合のネットワークプロセッサ11cは、新たなリクエストが同じファイルについて

のものであると、キャッシュメモリ11f(あるいは、メモリ11d)にキャッシュしておいたリプライメッセージをクライアント2へ返す応答部116(図2参照)としての機能も有していることになる。

【0021】このようにして、アクセス頻度の高い情報(メタ情報、ファイルデータ)に関しては、リダイレクタ11側でキャッシュして、NFSサーバ12-iにリクエストを転送することなくリダイレクタ11で応答を返してしまえば、アクセス頻度が高い情報に対するクライアント2への応答速度が大幅に向上し、本システム1としての処理速度及び処理能力が飛躍的に向上することになる。

【0022】次に、上記のNFSサーバ12-iは、それぞれ、リダイレクタ11から転送されてきたリクエストに応じたファイル処理(書き込み/更新/読み出しなど)を、内部ネットワーク(内部ネットワークスイッチ)14経由で二次記憶装置17にアクセスして実施したり、そのファイル処理結果をリクエスト元のクライアント2への応答として送付するためのリプライメッセージを生成してリダイレクタ11に送信したりすることができるものである。

【0023】なお、NFSサーバ12-iは、いずれも、ハードウェア的には、例えば図5に示すように、CPU(Central Processing Unit)12a、メモリ(主記憶部)12b及び内部ネットワーク14とのインタフェース(プロトコル変換など)を装備するインタフェースカード(IB-IF)12cをそなえて構成され、メモリ12bに記憶されたNFSサーバソフトウェア(プログラム)をCPU12aが読み取って動作することにより、上述したNFSサーバ12-iとしての機能が実現されるようになっている。

【0024】さて、ここで、これらのNFSサーバ12-iが、それぞれ独自にファイル名を管理すると、同じファイルデータ実体であるにも関わらず異なるNFSサーバ12-iで管理ファイル名が異なったり、逆に、異なるファイルデータ実体であるにも関わらず異なるNFSサーバ12-iで同じ管理ファイル名になったりといった不都合が生じ、NFSサーバ12-i間でファイルアクセス競合が生じる可能性がある。

【0025】このような不都合を解決するのが上記のネームサーバ13である。つまり、このネームサーバ13において、NFSサーバ12-iからのメタ情報アクセスを一元管理することで、全てのNFSサーバ12-iにおける管理ファイル名空間を1つにして、NFSサーバ12-i間でのファイルアクセス競合を回避するのである。従って、このネームサーバ13をそなえることにより、本システム1によるファイル共有の信頼性が大幅に向上することとなる。

【0026】なお、図1中に示すように、本ネームサーバ13には、故障などの異常発生(ダウン)にそなえて

現用と予備用とが存在する。また、これらのネームサーバ13についても、ハードウェア的には、NFSサーバ12-iと同様の構成(図5参照)、即ち、CPU13a、メモリ(主記憶部)13b及び内部ネットワーク14とのインタフェースを装備するインタフェースカード13cが実装されており、この場合も、メモリ13bに記憶されたネームサーバソフトウェア(プログラム)をCPU13aが読み取って動作することにより、上述したネームサーバ13としての機能が実現されるようになっている。

【0027】次に、上記の共有メモリ15は、上記のリダイレクタ11、NFSサーバ12-i及びネームサーバ13がそれぞれ内部ネットワーク14を介してアクセス可能なメモリで、例えば、或るNFSサーバ12-iあるいは現用のネームサーバ13がダウンした場合(障害発生時)に、そのNFSサーバ12-iあるいは現用のネームサーバ13の処理を他のNFSサーバ12-k($k=1\sim n$ で $k\neq i$)あるいは予備用のネームサーバ13に引継ぐための情報(以下、引継ぎ情報)などがサーバ12-i、13別にメモリカード(Shared Memory Card)15-1~15-m(mは自然数)に保持(バックアップ)されるようになっている(図6及び図8参照)。

【0028】つまり、上記の各NFSサーバ12-i(CPU12a)もしくはネームサーバ13(CPU13a)は、異常発生時にそなえて他のNFSサーバ12-iもしくは予備用のネームサーバ13に処理を引継ぐのに必要な情報を引継ぎ情報として共有メモリ15に記録する引継ぎ情報記録部121(131)(図5参照)としての機能を有していることになる。

【0029】なお、NFSサーバ12-iのダウンは、例えば図6に模式的に示すように、現用のネームサーバ13(CPU13a)がNFSサーバ監視デーモン132をバックグラウンドで実行することで監視し、現用のネームサーバ13のダウンは、例えば図8に模式的に示すように、予備用のネームサーバ13(CPU13a)がネームサーバ監視デーモン133をバックグラウンドで実行することで監視する。

【0030】そして、図7に模式的に示すように、NFSサーバ12-iのダウンが検出された場合(ステップS1)は、現用のネームサーバ13(CPU13a)が、ダウンしたNFSサーバ12-i以外のNFSサーバ12-k(例えば、負荷の軽いNFSサーバ12-k)に対して、ダウンしたNFSサーバ12-iの処理を引継ぐよう指示するとともに、リダイレクタ11に対してNFSサーバ12-iのダウンを通知する(ステップS2)。

【0031】これにより、引継ぎ指示を受けたNFSサーバ12-k(CPU12a)は、共有メモリ15に内部ネットワーク14を介してアクセスして、そこにバッ

クアップされている引継ぎ情報を読み出してダウンしたNFSサーバ12-iの処理を引継ぐ(ステップS3)。一方、このとき、リダイレクタ11(ネットワークプロセッサ11c)は、ネームサーバ13から上記の通知を受けることにより、ダウンしたNFSサーバ12-iにはリクエスト転送部112によるリクエストの転送を行なわないようにする。

【0032】つまり、この場合のネームサーバ13(CPU13a)は、NFSサーバ12-iの異常発生を検出する異常検出部134(図6参照)と、この異常検出部134でNFSサーバ12-iの異常発生が検出されると、そのNFSサーバ12-i以外の他のNFSサーバ12-kに対して異常発生したNFSサーバ12-iの処理を共有メモリ15の引継ぎ情報に基づいて引継ぐよう引継ぎ指示を与える引継ぎ指示部135(図6参照)としての機能を有していることになる。

【0033】このように、本実施形態では、たとえ、NFSサーバ12-iやネームサーバ13がダウンしたとしても、他のNFSサーバ12-kや予備用のネームサーバ13が処理を引継ぐことができるので、ストレージシステム1としては正常なファイル処理を継続することができ、耐障害性が大幅に向上する。なお、本例は、NFSサーバ12-iやネームサーバ13の冗長化についての説明であるが、勿論、リダイレクタ11を同様にして冗長化することも可能である。また、現用のネームサーバ13ダウン時の引継ぎについては、場合によってはNFSサーバ12-iのいずれかに引継ぐようにしてもよい。

【0034】次に、上述したリダイレクタ11からNFSサーバ12-iへのリクエスト転送時のリダイレクタ11及びNFSサーバ12-iでの具体的な処理について説明する。リダイレクタ11は、例えば、クライアント2から或るファイルデータを書き込むためのファイルアクセスリクエストを受信すると、そのファイルアクセスリクエストをリクエスト解析部111にて解析する。

【0035】なお、上記の「ファイルアクセスリクエスト」は、例えば図12に示すように、物理レイヤヘッダ(Phy Header)21aやIPヘッダ(Internet Protocol Header)21b、TCPヘッダ(Transmission Control Protocol Header)21c、NFSヘッダ21dなどから成るヘッダ部21と、実際に二次記憶装置17に書き込むべきファイルデータの実体(実ファイルデータ)が格納された実ファイルデータ部22とを有して成る。

【0036】そして、リクエスト解析部111は、図9に模式的に示すように、上記のファイルアクセスリクエスト中の実ファイルデータが始まる位置、即ち、ヘッダ部21と実ファイルデータ部22との境界をヘッダオフセット値[境界情報;例えば先頭からのビット数("a")]23として求める。求められた境界情報23は、リクエスト転送部112に通知され、リクエ

スト転送部112は、上記ファイルアクセスリクエストに通知された境界情報23を添付（付加）して転送先のNFSサーバ12-iに送付する。

【0037】つまり、上記のリクエスト解析部111は、図2中に示すように、受信したファイルアクセスリクエストを解析してそのファイルアクセスリクエストのヘッダ部21と実ファイルデータ部22との境界位置を表わすヘッダオフセット値23を求めるヘッダオフセット値解析部111aとしての機能と、このヘッダオフセット値解析部111aで得られたヘッダオフセット値23をNFSサーバ12-iへ転送されるファイルアクセスリクエストに付加するヘッダオフセット値付加部111bとしての機能とを有していることになる。

【0038】その後、NFSサーバ12-i側では、上述のごとくりダイレクタ11側で付加されたヘッダオフセット値23に基づいて、NIC（Network Interface Card）ドライバ（ネットワークドライバ）122が、実ファイルデータ部22とそれ以外の領域（ヘッダ部21）の先頭アドレスをそれぞれ上位層（NFS処理層）のカーネル内（カーネル上位層）で扱われるメッセージのページ境界（ページ境界（別領域）；バッファ（mbuf）123、124）に割り当てることができる（図10参照）。

【0039】このようにすると、例えば図11に模式的に示すように、ファイルアクセスリクエストがカーネル上位層のファイルシステム部125に到達したときに、実ファイルデータ部22の先頭アドレス（ポインタ）をファイルシステムバッファ126へのポインタに付け替えるだけで、データのコピーを発生させずに（マップ切り替え）、ファイルシステムバッファ126にデータを移すことが可能になる（カーネル内ゼロコピーが実現される）。従って、DMA（Direct Memory Access）も高速に実行でき、NFSサーバ12-iの処理速度及び処理能力を大幅に向上することができる。

【0040】なお、ヘッダ部21と実ファイルデータ部22との境界は、NICドライバ122側で求めるようにすることも考えられるが、この場合、NICドライバ122の処理（ヘッダ解析）量が増大する（通常、NICドライバ122は物理レイヤヘッダ21aの解析のみを行なう）ため、上記のように元々ヘッダ部21の解析機能（リクエスト解析部111）を有するリダイレクタ11側で境界を求めるようにした方が、NICドライバ層での処理量を増やさずに（処理能力低下を招かず）上位層（NFS処理層）でのカーネルゼロコピーを実現できる。

【0041】以上のように、本実施形態のストレージシステム1によれば、システム1内部に、リダイレクタ11、複数のNFSサーバ12-i、ネームサーバ13、共有メモリ15、二次記憶装置17を設け、これらを高速な内部ネットワーク14で接続する構成にしたことに

より、必要に応じてNFSサーバ12-iや二次記憶装置17を簡単に増設することができ、しかも、NFSサーバ12-iの維持（保守）を個々に行なう必要もないので、外部ネットワーク3の帯域拡大に対して低コストで十分に対応できる（例えば、10Gbps LANまで対応可能な）性能・容量スケーラビリティを確保することができる。

【0042】特に、上述した実施形態では、リダイレクタ11が、各NFSサーバ12-iの負荷に応じて各NFSサーバ12-iに均一に処理が割り当てられるように制御したり、同じファイルについてのリクエストに対する処理は同じNFSサーバ12-iに割り当てたり、アクセス頻度の高いファイルについてのリクエストに対してはNFSサーバ12-i側ではなくリダイレクタ11側でキャッシュしておいたリプライメッセージにより応答したりするので、その処理速度及び性能が飛躍的に向上しており、10Gbps LANまで確実に対応可能な性能・容量スケーラビリティが実現されている。

【0043】（B）第1変形例の説明

上述したシステム1内には、メモリ12bの容量が他のNFSサーバ12-iよりも大きいNFSサーバ12-iをキャッシュサーバ12'（図1参照）として配置して、このキャッシュサーバ12'でのファイルアクセスは基本的にメモリ12bに対する読み書きのみでクライアント2側に応答を返すようにしてもよい。

【0044】そして、一定期間内のリクエスト頻度が所定回数（閾値）以上であるアクセス頻度の高いファイルについてはキャッシュサーバ12'のメモリ12bにキャッシュしておき、キャッシュサーバ12'が応答するようにする。具体的には、まず、リダイレクタ11側で各ファイルのアクセス頻度を監視しておき、アクセス頻度が或る閾値よりも高いファイルへのアクセスに関しては、キャッシュサーバ12-iで処理を行なうように、リダイレクタ12-iからネームサーバ13、NFSサーバ12-i、キャッシュサーバ12'に指示を与える。

【0045】つまり、このとき、リダイレクタ11（リクエスト転送部112）は、アクセス頻度の高いファイルについてのリクエストをキャッシュサーバ12'に転送することになる。これにより、アクセス頻度の高いファイルに対するアクセスは、二次記憶装置17にアクセスすることなくキャッシュサーバ12'内で処理されるので、ストレージシステム1としての処理速度及び処理能力の大幅な向上に大きく寄与する。

【0046】一方、上記のアクセス頻度の高いファイルに対するアクセス頻度が落ちてくると（キャッシュサーバ12'のメモリ12bにキャッシュされているファイルに対するアクセス頻度が所定回数以下になると）、リダイレクタ11が適当な（例えば、負荷の軽い）NFSサーバ12-iを割り当てて処理を移行するようにネー

ムサーバ13、NFSサーバ12-i、キャッシュサーバ12'に指示する。

【0047】つまり、この場合、リダイレクタ11（リクエスト転送部112）は、リクエストの転送先をキャッシュサーバ12'以外のNFSサーバ12-iに変更するのである。これにより、アクセス頻度が落ちてきたファイルがいつまでもキャッシュサーバ12'にキャッシュされ続けることが回避され、その結果、キャッシュサーバ12'に必要なメモリ容量を削減することができるのと同時に、キャッシュサーバ12'での処理に余裕をもたせてその処理能力を向上することができる。

【0048】（C）第2変形例の説明

なお、例えば図13に示すように、上記の二次記憶装置17をFCスイッチ18経由でネームサーバ13及びNFSサーバ12-iと接続（二次記憶ネットワークを構築）し、FCスイッチ18と外部ノード19とを接続することにより、二次記憶装置17に外部ノード19からもアクセスできるようにすることも可能である。ただし、この場合、外部ノード19で動作させるファイルシステムは、ストレージシステム1内のファイルシステムと同じもの（上述した例では、NFS；図13中の網かけ部がこれを意味する）である必要がある。

【0049】このようにすることで、システム1内部のNFSサーバ12-iとのアクセス競合を避けるために外部ノード19からのアクセスについては何らかの調停制御が必要になるが、外部ノード19から本ストレージシステム1内のファイルへのアクセスが可能になる。ただし、例えば図14に示すように、外部ノード19からネームサーバ13へのアクセスを許容するようにすれば、外部ノード19からのアクセスについてもシステム1内の管理ファイル名に従うことになるので、上記の調停制御を必要とせず外部ノード19からのファイルアクセスが行なえるようになる。なお、この図14では、内部ネットワーク14経由でのネームサーバ13へのアクセスを許容する場合であるが、勿論、図13において二次記憶ネットワーク（FCスイッチ18）経由でのアクセスを許容するようにしてもよい。

【0050】また、例えば図15に示すように、NFSサーバ12-iと外部ノード19とを共通化してしまってもよい。つまり、NFSサーバ12-iを、外部ネットワーク3から直接受信されるリクエストに応じたファイル処理を二次記憶装置17に対して行なうように構成するのである。これにより、或るクライアント2がリダイレクタ11経由でアクセスしてきた場合は、NFSサーバ12-iは上述したストレージシステム1のファイルサーバとして機能し、直接、NFSサーバ12-iにアクセスしてきた場合は、リダイレクタ11を経由せずに応答する通常のファイルサーバとして機能することになる。つまり、クライアント2からのNFSサーバ12-i経由でのアクセスとNFSサーバ12-iを経由し

ない直接アクセスとの双方を許容するのである。

【0051】いずれの場合も、外部からの直接アクセスが可能となり、他のストレージアーキテクチャ（例えば、SAN（Storage Area Network）など）との融合を実現することができる。なお、図14及び図15において、符号20は内部ネットワーク14と二次記憶装置17とのインタフェースを装備するネットワークディスクアダプタを表す。

【0052】また、図13～図15に示す構成では、前述した共有メモリ15が省略されているが、勿論、装備されていてもよい。このようにすれば、図13～図15に示す構成においても、前記と同様のバックアップ処理が可能になる。

（D）その他

なお、上述した実施形態では、内部ネットワーク14としてInfiniband、外部ネットワーク3としてギガビットイーサネットを適用した場合について説明したが、勿論、これら以外の他の高速ネットワークを用いてシステム構築することも可能である。

【0053】また、上記のネームサーバ13や共有メモリ15は必ずしもそなえる必要はなく、これらのいずれか、あるいは、双方を省略しても本発明の目的は十分達成される。さらに、上述した実施形態では、ファイルサーバにNFSを適用しているが、本発明はこれに限定されず、勿論、他のファイルシステムを適用することも可能である。

【0054】また、上述した実施形態では、内部ネットワーク14の容量（通信速度）が4～10Gbps程度であることを前提としたが、この速度は外部ネットワーク3の帯域拡大に応じて適宜変更すれば対応できる。そして、本発明は、上述した実施形態に限定されず、上記以外にも本発明の趣旨を逸脱しない範囲で種々変形して実施することができる。

【0055】（E）付記

【付記1】 ファイルデータを記憶しうる記憶装置と、リクエストに応じたファイル処理を該記憶装置に対して行なう複数のファイルサーバと、外部ネットワークを介してクライアントから受信されるリクエストの該ファイルサーバへの転送処理と、該リクエストに対する該クライアントへの応答処理とを一元管理するファイルサーバ管理ノードと、該記憶装置、該ファイルサーバ及び該ファイルサーバ管理ノードを通信可能に相互接続する内部ネットワークとをそなえて構成されたことを特徴とする、ストレージシステム。

【0056】【付記2】 該内部ネットワークに、該ファイルサーバが扱うファイルデータ名を一元管理するネームサーバが接続されていることを特徴とする、付記1記載のストレージシステム。

【付記3】 該内部ネットワークに、該ファイルサーバ管理ノード及び該ファイルサーバがアクセス可能な共有

メモリが接続されていることを特徴とする、付記1記載のストレージシステム。

【0057】〔付記4〕 該内部ネットワークに、該ファイルサーバ管理ノード、該ファイルサーバ及び該ネームサーバがアクセス可能な共有メモリが接続されていることを特徴とする、付記2記載のストレージシステム。

〔付記5〕 該ファイルサーバ管理ノードが、該リクエストの内容を解析するリクエスト解析部と、該リクエスト解析部の解析結果に応じて該リクエストを特定のファイルサーバに転送するリクエスト転送部とをそなえていることを特徴とする、付記1～4のいずれか1項に記載のストレージシステム。

【0058】〔付記6〕 該ファイルサーバ管理ノードが、該リクエスト転送部による過去のリクエストの転送履歴を記録する転送履歴記録部をそなえ、該リクエスト転送部が、該転送履歴記録部の該転送履歴に基づいて同一ファイルデータ名のファイルデータに対するリクエストを同一ファイルサーバに転送するように構成されたことを特徴とする、付記5記載のストレージシステム。

【0059】〔付記7〕 該ファイルサーバ管理ノードが、該ファイルサーバの負荷を監視する負荷監視部をそなえ、該リクエスト転送部が、該負荷監視部での監視結果に基づいて負荷の低いファイルサーバに該リクエストを転送するように構成されたことを特徴とする、付記5記載のストレージシステム。

【0060】〔付記8〕 該ファイルサーバのうち少なくとも1台が、該記憶装置におけるファイルデータをキャッシュしうる主記憶部をそなえ、該主記憶部において該リクエストに応じたファイル処理を実行するキャッシュサーバとして構成されていることを特徴とする、付記5記載のストレージシステム。

〔付記9〕 該キャッシュサーバの該主記憶部が、一定期間内のリクエスト頻度が所定回数以上のファイルデータをキャッシュしておくように構成されるとともに、該リクエスト転送部が、上記のリクエスト頻度の高いファイルデータについてのリクエストを該キャッシュサーバに転送するように構成されたことを特徴とする、付記8記載のストレージシステム。

【0061】〔付記10〕 該リクエスト転送部が、該キャッシュサーバの該主記憶部にキャッシュされている該ファイルデータに対するリクエスト頻度が所定回数以下になると、該リクエストの転送先を該キャッシュサーバ以外のファイルサーバに変更するように構成されたことを特徴とする、付記9記載のストレージシステム。

【0062】〔付記11〕 該リクエスト解析部が、該リクエストを解析して当該リクエストのヘッダ部と実ファイルデータ部との境界位置を表わすヘッダオフセット値を求めるヘッダオフセット値解析部と、該ヘッダオフセット値解析部で得られた該ヘッダオフセット値を該ファイルサーバへ転送される該リクエストに付加するヘッ

ダオフセット値付加部とをそなえていることを特徴とする、付記5記載のストレージシステム。

【0063】〔付記12〕 該ファイルサーバが、該リクエストに付加された該ヘッダオフセット値に基づいて該リクエストの該ヘッダ部と該データ部とをそれぞれカーネル上位層で扱われるメッセージの異なる領域にコピーするネットワークドライバをそなえていることを特徴とする、付記11記載のストレージシステム。

〔付記13〕 該ファイルサーバ管理ノードが、リクエスト頻度の高い特定のファイルデータについての該クライアントへの応答メッセージをキャッシュしておくキャッシュ部と、該リクエストが該特定のファイルデータについてのものであると、該キャッシュ部の該当該応答メッセージを該クライアントへ返す応答部とをそなえていることを特徴とする、付記1～12のいずれか1項に記載のストレージシステム。

【0064】〔付記14〕 該ファイルサーバが、異常発生時にそなえて他のファイルサーバに処理を引継ぐのに必要な情報を引継ぎ情報として該共有メモリに記録する引継ぎ情報記録部をそなえていることを特徴とする、付記3又は付記4に記載のストレージシステム。

〔付記15〕 該ファイルサーバの異常発生を検出する異常検出部と、該異常検出部で該ファイルサーバの異常発生が検出されると、当該ファイルサーバ（以下、異常ファイルサーバという）以外の他のファイルサーバに対して該異常ファイルサーバの処理を該共有メモリの該引継ぎ情報に基づいて引継ぐよう引継ぎ指示を与える引継ぎ指示部とが設けられたことを特徴とする、付記14記載のストレージシステム。

【0065】〔付記16〕 該記憶装置が、外部ノードからのアクセスを許容するように構成されたことを特徴とする、付記1～15のいずれか1項に記載のストレージシステム。

〔付記17〕 該ネームサーバが、外部ノードからのアクセスを許容するように構成されたことを特徴とする、付記2～15のいずれか1項に記載のストレージシステム。

【0066】〔付記18〕 該ファイルサーバが、該外部ネットワークから直接受信されるリクエストに応じたファイル処理を該記憶装置に対して行なうように構成されたことを特徴とする、付記1～17のいずれか1項に記載のストレージシステム。

【0067】

【発明の効果】以上詳述したように、本発明のストレージシステムによれば、リクエストに応じたファイル処理を記憶装置に対して行なう複数のファイルサーバと、各ファイルサーバの処理を一元管理するファイルサーバ管理ノードと、記憶装置、ファイルサーバ及びファイルサーバ管理ノードを通信可能に相互接続する内部ネットワークとをそなえているので、必要に応じてファイルサー

バや記憶装置を簡単に増設することができ、しかも、各ファイルサーバの維持（保守）を個々に行なう必要もないので、外部ネットワークの帯域拡大に対して低コストで十分に対応できる性能・容量スケーラビリティを確保することができる。

【0068】そして、上記の内部ネットワークに、上記の各ファイルサーバが扱うファイルデータ名を一元管理するネームサーバを接続すれば、ファイルサーバ間でファイルアクセス競合が生じることを防止することができるので、ファイル共有の信頼性向上に大きく寄与する。また、上記の内部ネットワークに、共有メモリを接続して、この共有メモリに障害発生時にそなえたファイルサーバの引継ぎ情報を随時記憶するようにしておくことで、一部のファイルサーバに障害が発生したとしても、ストレージシステムとしては正常な処理を継続することができるので、耐障害性が大幅に向上する。

【0069】さらに、ファイルサーバ管理ノードは、過去のリクエストの転送履歴に基づいて同一ファイルデータ名のファイルデータに対するリクエストを同一ファイルサーバに転送するように構成してもよく、このようにすれば、処理速度が大幅に向上する。また、ファイルサーバの負荷を監視して負荷の低いファイルサーバにリクエストを転送するようにしてもよく、このようにすれば、各ファイルサーバに均一にリクエストを転送できるので、一部のファイルサーバに負荷が集中して障害を引き起こしてしまうようなことを確実に防止することができる、信頼性が大幅に向上する。

【0070】さらに、リクエスト頻度が高いファイルデータについては、キャッシュサーバの主記憶部にキャッシュしておき、キャッシュサーバで処理するようにしておけば、記憶装置へのアクセス頻度を大幅に削減することができるので、さらに処理速度及び処理性能が向上する。そして、この場合、キャッシュサーバの主記憶部にキャッシュされているファイルデータに対するリクエスト頻度が所定回数以下になると、キャッシュサーバ以外のファイルサーバで処理を行なうようにすれば、リクエスト頻度の低いファイルデータがいつまでもキャッシュサーバに保持され続けることが回避されるので、キャッシュサーバの主記憶部に必要なメモリ容量を削減することができるとともに、キャッシュサーバでの処理に余裕をもたせてその処理能力を向上することができる。

【0071】また、ファイルサーバ管理ノードにおいて、リクエストのヘッダ部と実ファイルデータ部との境界位置を表わすヘッダオフセット値を求めて、そのヘッダオフセット値をリクエストに付加してファイルサーバに転送するようにすれば、ファイルサーバのネットワークドライバにおいて、そのヘッダオフセット値に基づいてリクエストのヘッダ部とデータ部とをそれぞれカーネル上位層で扱われるメッセージの異なる領域にコピーすることができる。従って、カーネル内ゼロコピーを実現

することができ、ファイルサーバの処理速度及び処理能力を大幅に向上することができる。

【0072】さらに、上記のファイルサーバ管理ノードにおいて、リクエスト頻度の高い特定のファイルデータについてのクライアントへの応答メッセージをキャッシュしておき、受信したリクエストがそのファイルデータについてのものであると、キャッシュしておいた応答メッセージをクライアントへ返すようにすれば、ファイルサーバへリクエストを転送する必要が無いので、クライアントに対する応答速度が大幅に向上し、本システムとしての処理速度及び処理能力が飛躍的に向上することになる。

【0073】また、上記の記憶装置は、外部ノードからのアクセスを許容するように構成してもよく、このようにすれば、他のストレージアーキテクチャとの融合が可能になる。ここで、外部ノードから上記のネームサーバへのアクセスを許容するようにすれば、上記のファイルサーバと外部ノードとのファイルアクセス調停制御を必要とせず、外部ノードからのファイルアクセスが可能になる。

【0074】さらに、上記のファイルサーバは、上記の外部ネットワークから直接受信されるリクエストに応じたファイル処理を記憶装置に対して行なうように構成してもよく、このようにすれば、クライアントからのファイルサーバ経由でのアクセスとファイルサーバを経由しない直接アクセスとの双方を許容することができるので、この場合も、他のストレージアーキテクチャとの融合が可能になる。

【図面の簡単な説明】

【図1】本発明の一実施形態としてのストレージシステムの構成（ストレージアーキテクチャ）を示すブロック図である。

【図2】図1に示すリダイレクタの構成を示すブロック図である。

【図3】図1に示すリダイレクタでメタ情報をキャッシュする場合を説明するためのブロック図である。

【図4】図1に示すリダイレクタでリプライメッセージを返す場合を説明するためのブロック図である。

【図5】図1に示すNFSサーバ（ネームサーバ）の構成を示すブロック図である。

【図6】図1に示す共有メモリにNFSサーバの引継ぎ情報をバックアップする場合を説明するためのブロック図である。

【図7】図6に示す共有メモリにバックアップされた引継ぎ情報に基づいてダウンしたNFSサーバの処理を引継ぐ場合を説明するためのブロック図である。

【図8】図1に示す共有メモリにネームサーバの引継ぎ情報をバックアップする場合を説明するためのブロック図である。

【図9】図1に示すリダイレクタにおいてファイルアク

セスリクエストの境界情報を求める場合を説明するためのブロック図である。

【図10】図1に示すNFSサーバにおいて図9に示すファイルアクセスリクエストの境界情報に基づいてカーネル内ゼロコピーを実現する場合を説明するためのブロック図である。

【図11】図1に示すNFSサーバにおいて図9に示すファイルアクセスリクエストの境界情報に基づいてカーネル内ゼロコピーを実現する場合を説明するためのブロック図である。

【図12】図9～図11に示すファイルアクセスリクエストのフォーマット例を示す図である。

【図13】図1に示すストレージシステムにおいて外部ノードからの二次記憶装置へのアクセスを許容する場合の構成を示すブロック図である。

【図14】図1に示すストレージシステムにおいて外部ノードからのネームサーバへのアクセスを許容する場合の構成を示すブロック図である。

【図15】図1に示すストレージシステムにおいて外部ノードからのリダイレクタ経由のアクセスと直接アクセスとを許容する場合の構成示すブロック図である。

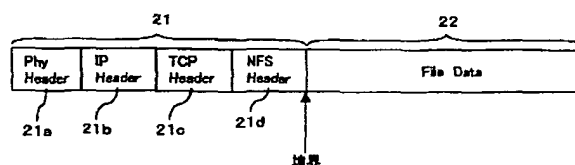
【図16】ネットワーク上での複数ノード（クライアント）間のファイル共有を実現する従来の手法を説明するためのブロック図である。

【符号の説明】

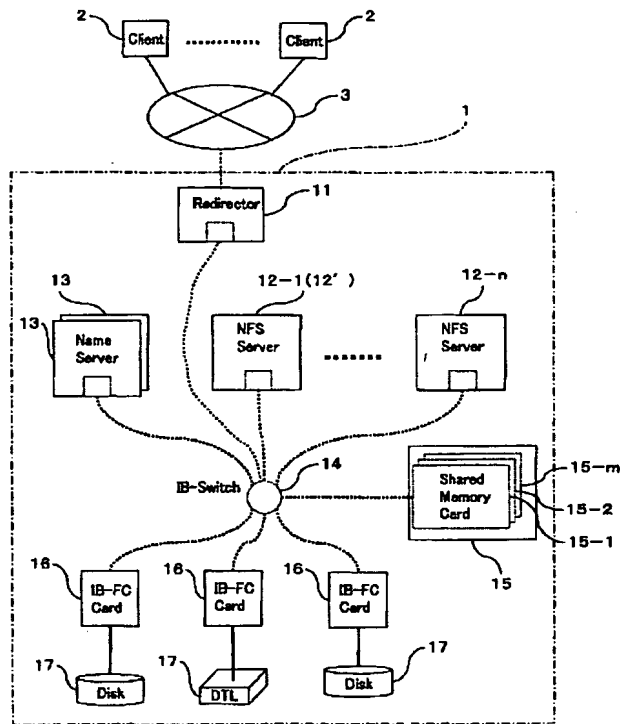
- 1 ストレージシステム
- 2 クライアント
- 3 外部ネットワーク（ギガビットイーサネット）
- 11 リダイレクタ（Redirector；ファイルサーバ管理ノード）
- 11a ギガビットイーサネットカード
- 11b インフィニバンドカード
- 11c ネットワークプロセッサ
- 11d メモリ（主記憶部）
- 11e PCI（Peripheral Component Interconnect）バス
- 11f キャッシュメモリ（キャッシュ部）
- 12-1～12-n NFS（Network File System）サーバ（ファイルサーバ）
- 12' キャッシュサーバ

- 12a, 13a CPU（Central Processing Unit）
- 12b, 13b メモリ（主記憶部）
- 12c, 13c インタフェースカード（IB-IF）
- 13 ネームサーバ
- 14 高速（内部）ネットワーク〔インフィニバンド（Infiniband）〕スイッチ
- 15 共有メモリ（Shared Memory）
- 15-1～15-m メモリカード（Shared Memory Card）
- 16 IB-FCカード
- 17 二次記憶装置
- 18 FCスイッチ
- 19 外部ノード
- 20 ネットワークディスクアダプタ
- 21 ヘッダ部
- 21a 物理レイヤヘッダ（Phy Header）
- 21b IPヘッダ（Internet Protocol Header）
- 21c TCPヘッダ（Transmission Control Protocol Header）
- 21d NFSヘッダ
- 22 実ファイルデータ部
- 23 ヘッダオフセット値（境界情報）
- 111 リクエスト解析部
- 111a ヘッダオフセット値解析部
- 111b ヘッダオフセット値付加部
- 112 リクエスト転送部
- 113 転送履歴記録部
- 114 負荷監視部
- 115 NFSサーバ負荷監視デーモン（daemon）
- 116 応答部
- 121, 131 引継ぎ情報記録部
- 122 NIC（Network Interface Card）ドライバ（ネットワークドライバ）
- 123, 124 バッファ（mbuf）
- 125 ファイルシステム部
- 126 ファイルシステムバッファ
- 132 NFSサーバ監視デーモン
- 133 ネームサーバ監視デーモン
- 134 異常検出部
- 135 引継ぎ指示部

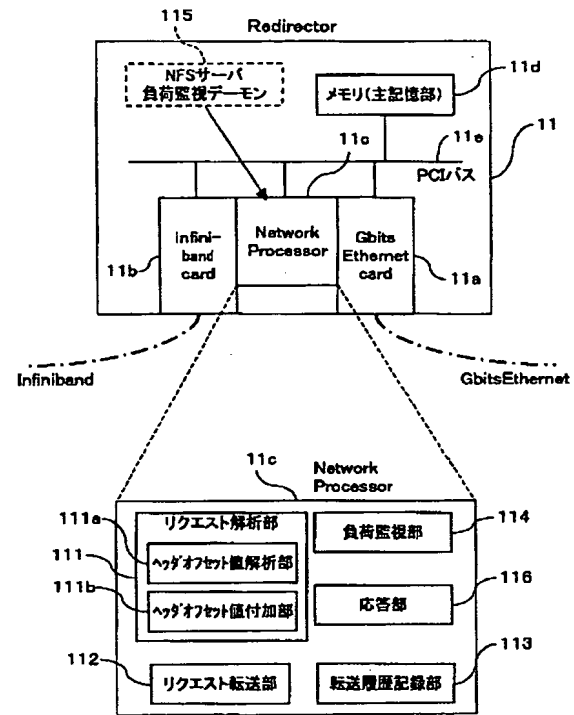
【図12】



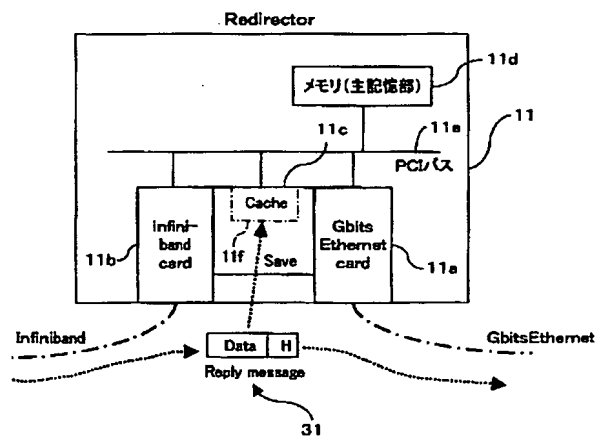
【図1】



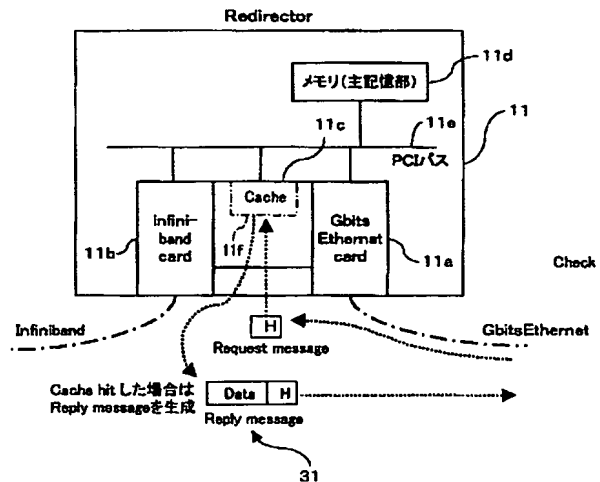
【図2】



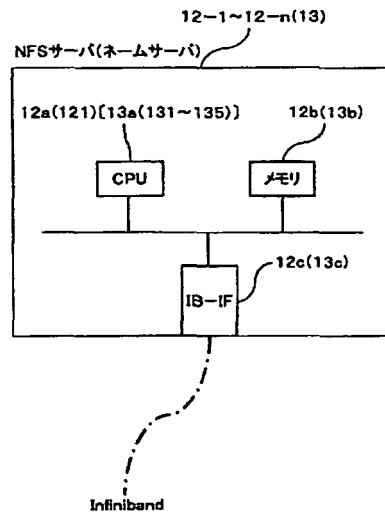
【図3】



【図4】

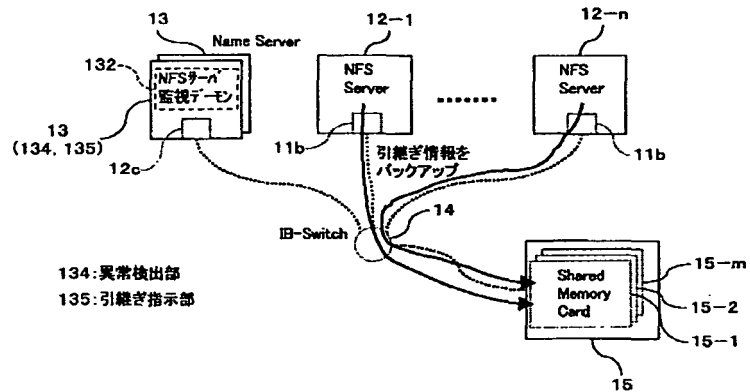


【図5】

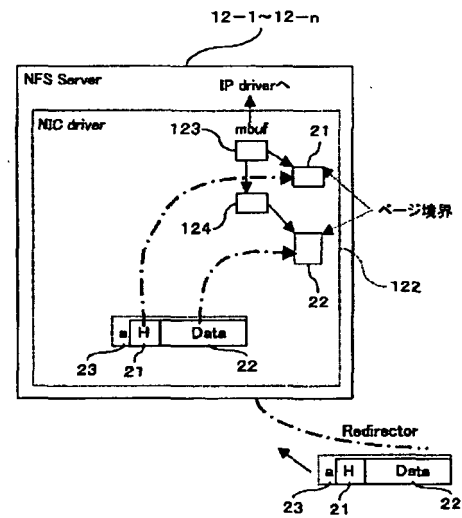


121, 131: 引継ぎ情報記録部
 132: NFSサーバ監視デモン
 133: ネームサーバ監視デモン
 134: 異常検出部
 135: 引継ぎ指示部

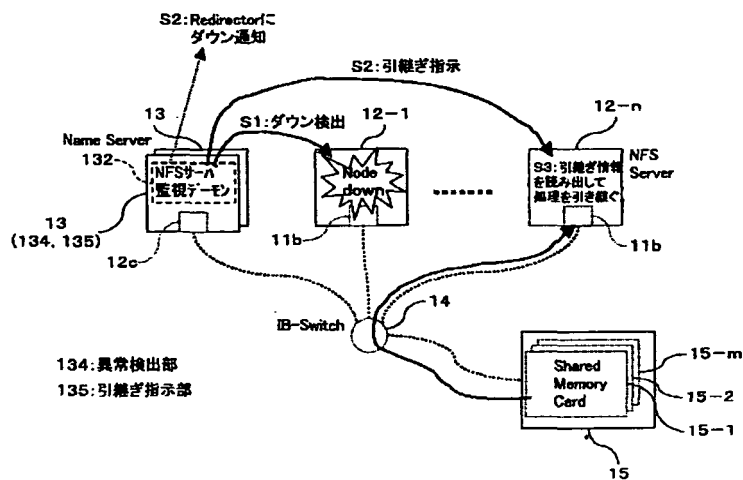
【図6】



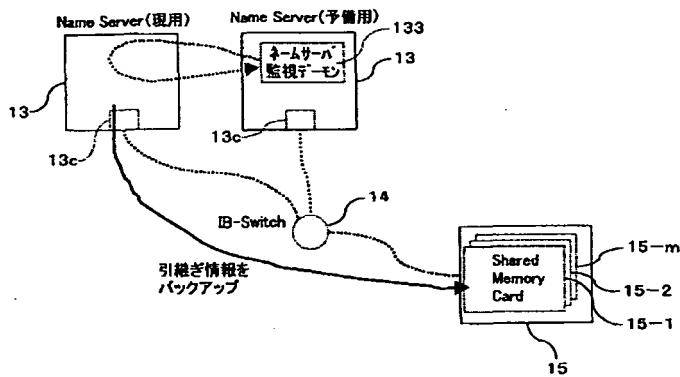
【図10】



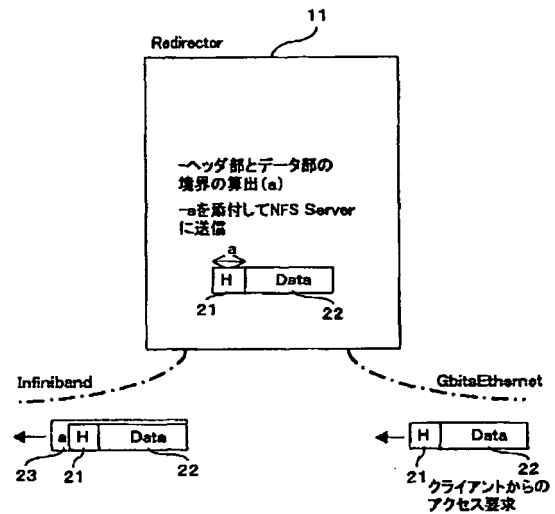
【図7】



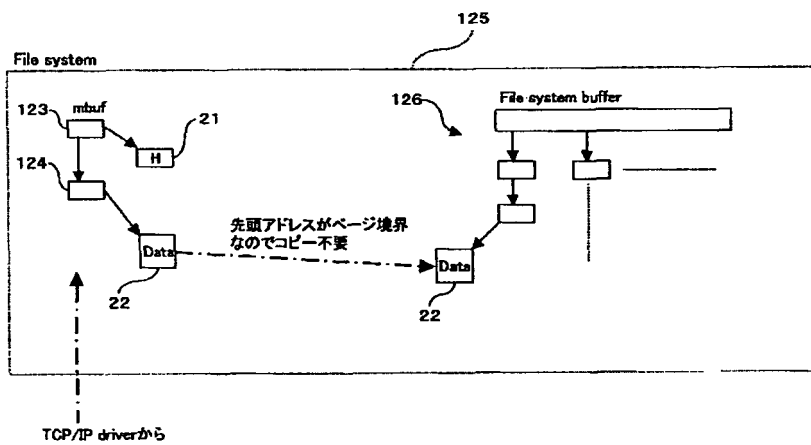
【図 8】



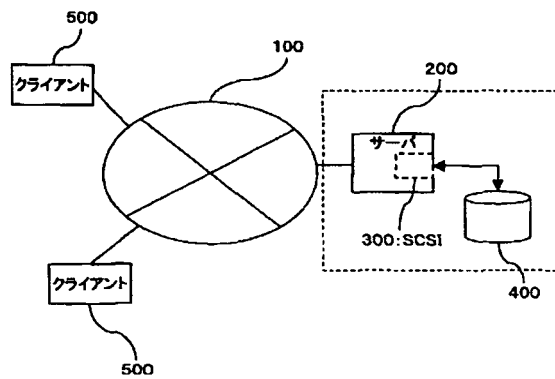
【図 9】



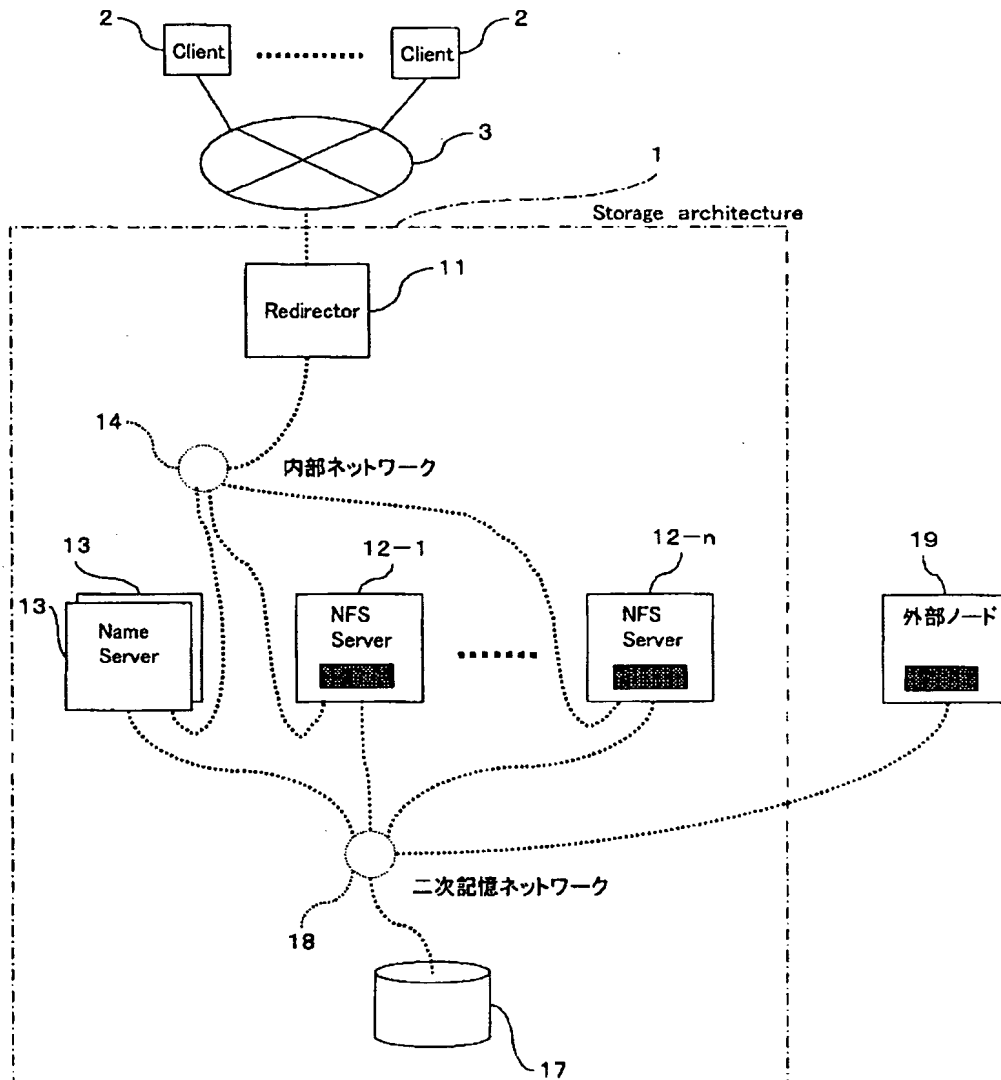
【図 11】



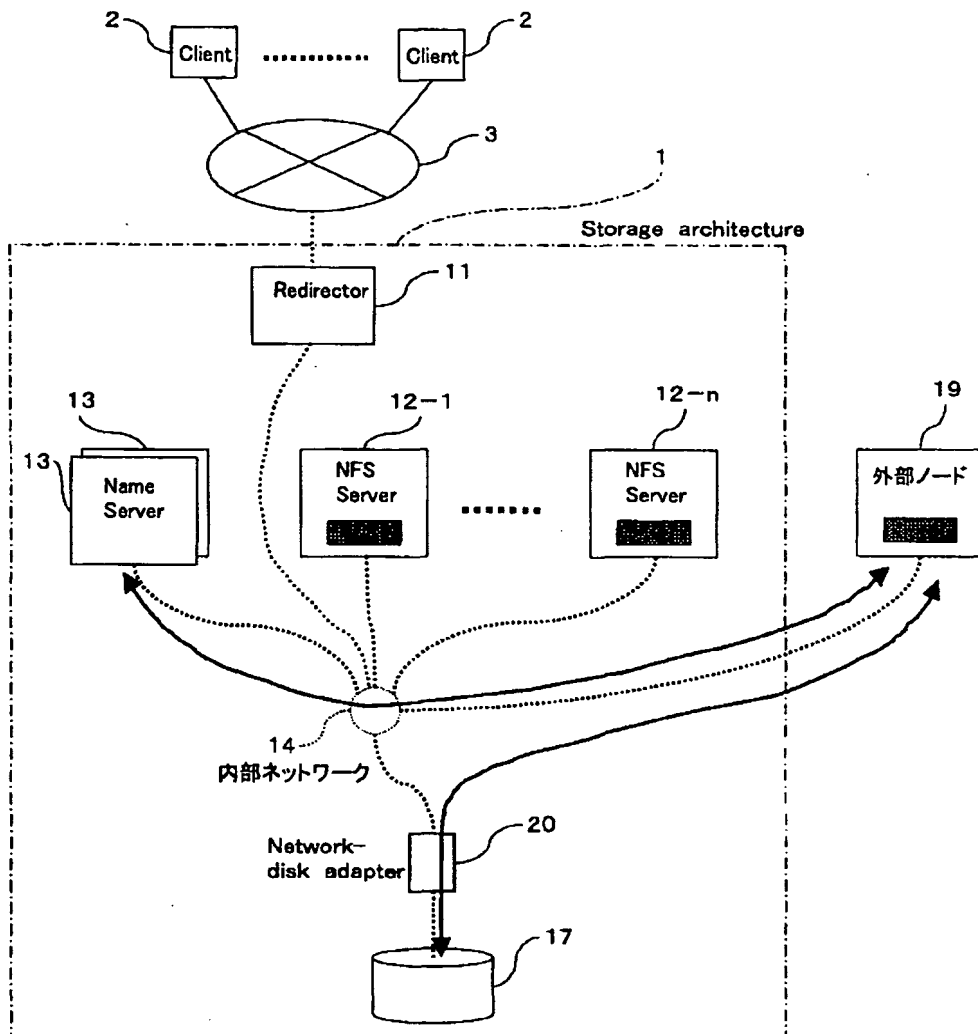
【図 16】



【図13】



【図14】



【図15】

